THE UNIVERSITY of EDINBURGH

Edinburgh Research Explorer

Strawson, moral responsibility, and the "Order of Explanation"

Citation for published version:
Todd, P 2016, 'Strawson, moral responsibility, and the "Order of Explanation": An intervention', *Ethics: An International Journal of Social, Political, and Legal Philosophy*, vol. 127, no. 1, pp. 208-240. https://doi.org/10.1086/687336

Digital Object Identifier (DOI):
10.1086/687336

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
Ethics: An International Journal of Social, Political, and Legal Philosophy

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

OPEN ACCESS

Download date: 06. May. 2023

Strawson, Moral Responsibility, and the "Order of Explanation": An Intervention*

Introduction

Perhaps no paper on the topic of moral responsibility in the last 60 years has had more influence – or has been more widely discussed – than P.F. Strawson's (1962) "Freedom and Resentment". The paper has provoked a wide ride range of responses, both positive and negative, and an equally wide range of different interpretations. In particular, beginning with Gary Watson, some have seen Strawson as making (or at least suggesting) a point about the "order of explanation" concerning moral responsibility. In some sense, allegedly, Strawson wished to "reverse" the traditional order of explanation: it is not that it is appropriate to hold agents responsible because they *are* morally responsible, rather, it is … well, something else. The "traditional" view has it that it is appropriate to *hold* agents responsible because they meet certain objective conditions on *being* responsible. Strawson, allegedly, wished to say something different, something incompatible with this "traditional" understanding. Strawsonian claims about the "facts of responsibility" and the "order of explanation" are often stated in different ways, but one thing remains constant in such presentations: either implicitly or explicitly, such proposals are drawn *in contrast* to libertarian theories of moral responsibility. According to those who discuss this "reversal", the reversal is (or is at least meant to be) incompatible with the views of those who (in Strawson's words) "over-intellectualize" the facts of moral responsibility – in particular, and paradigmatically, with the views of libertarians who say that a kind of freedom that is incompatible with determinism is required in order to be morally responsible.

The overarching theme of this paper is that extant developments of "the reversal" face a dilemma: in order to make the relevant proposals plausibly anti-libertarian, they must be made to be implausible on other grounds. Broadly speaking, the dilemma pertains to Strawsonian struggles with the word *appropriate*. Sometimes the relevant proposal appears to be that whether a given agent is morally responsible is determined by whether it is appropriate to *hold* her responsible. Even if such a conception of the "facts of moral responsibility" is plausible, however, it is not plausibly anti-libertarian: the libertarian can accept that a given agent is morally responsible because it is appropriate to hold her

responsible. On the other hand, sometimes the proposal instead seems to be, not that whether a given agent is responsible is determined by whether it is *appropriate* that she be held responsible, but instead determined by whether we actually *do* (or perhaps are *disposed* to) hold her responsible. But whereas such a conception is plausibly anti-libertarian, it is, it would seem, implausible on other grounds – grounds that are not often acknowledged by theorists developing such proposals. In particular, such theories would not seem to have the resources to explain why, even if we did (or were disposed to) blame the severely mentally ill and small children, they nevertheless would not be blameworthy.

My strategy in this paper is simply to consider and then critically evaluate representative statements of this "reversal" by various theorists of moral responsibility. As far as I can discover, the first explicit statement of this kind is due to Gary Watson. Similar statements to those of Watson can subsequently be found in a variety of discussions of moral responsibility, but I will consider in particular passages from Robert Kane, David Brink and Dana Nelkin, Justin Coates and Neal Tognazzini, and finally from Tognazzini (working alone). As I hope will become clear, none of these articulations is fit for the purposes for which it is intended. I then consider R.J. Wallace's important development of similar themes in his *Responsibility and the Moral Sentiments*. As we'll see, though Wallace does not articulate a "reversal" thesis in the sense at stake, his position does encounter problems similar to the other proposals I consider in this paper. Accordingly, if we are to understand the given "reversal", we need a new approach, and in the final section of the paper, I suggest an avenue by which one might be constructed: an analogy with the concept of *funniness*.

At the outset, I should be clear on how my paper is intended to relate to Strawson's original. I aim to stay neutral on the question whether any "reversal" thesis was intended by Strawson himself. Certainly Strawson nowhere makes any explicit point about "reversing" the "order of explanation" regarding moral responsibility – such a thesis is, at most, implicit in Strawson's essay. Accordingly, I take no stand on the question whether any such "reversal" is an essential part of Strawson's (or even a "Strawsonian") overall approach to moral responsibility. For instance, Wallace sees in Strawson "the seeds for at least three different kinds of argument against incompatibilism." (1994: 96) Perhaps only one such argument should be identified as at the core of Strawson's project, and perhaps that

argument does not rely on anything like a "reversal". These are difficult matters.[1] Thankfully, we need not resolve them before investigating the issues this paper seeks to address: those concerning the reversal itself. To such issues I now turn.

Watson

Watson most explicitly expresses the relevant claim about the "order or explanation" in a certain passage of his classic 1987 paper, "Responsibility and the Limits of Evil: Variations on a Strawsonian Theme". Broadly speaking, Watson here seeks to articulate the way in which Strawson's approach differs from the "consequentialist" approach and the "libertarian" approach to moral responsibility, both of which, according to Strawson, "over-intellectualize the facts". On the one hand, the consequentialist seeks to justify our "reactive attitudes" (such as blame and resentment) in terms of their social efficacy, whereas the libertarian instead seeks to justify such attitudes on grounds that the relevant parties possess a kind of freedom that is incompatible with determinism; according to Strawson, however, neither such "justification" is required. Let me begin, oddly, with what Watson did *not* say, but what it seems to me one would have *expected* Watson to say, given what comes immediately prior in Watson's text. Watson says (but for the final line):

> What these otherwise very different views [of the consequentialist and the libertarian] share is the assumption that our reactive attitudes commit us to the truth of some independently apprehensible proposition which gives the content of the belief in responsibility; and so either the search is on for the formulation of this proposition, or we must rest content with an intuition of its content. For the social-regulation theorist, this is a proposition about the standard effects of having and expressing

[1] Cf. Wallace 1994:10:

> The very complexity of Strawson's position…virtually ensured that its legacy would be similarly complex and multifaceted. Thus there is no fixed and stable view that might be labeled the Strawsonian account of responsibility.

For sustained treatments of Strawson's "project" that do not employ a "reversal" thesis, see, e.g., Russell 2011: 199 – 220, Pereboom 2001: 90 – 100, and Bennett's classic presentation in his 1980 "Accountability" (reprinted as his 2008: 47 – 68, though see pg. 55 for a point perhaps *similar* to that of the "reversal").

> reactive attitudes. For the libertarian, it is a proposition concerning metaphysical freedom. Since the truth of the former is consistent with the thesis of determinism, the consequentialist is a compatibilist; since the truth of the latter is shown or seen not to be, the libertarian is an incompatibilist.
>
> In Strawson's view, there is no such independent notion of moral responsibility that explains the propriety of the reactive attitudes. The explanatory priority is the other way around: It is not that we hold people responsible because they *are* responsible; rather, they are responsible because we hold them responsible.[2]

If this passage had ended in this way, then it is (relatively) clear where the philosophical action would be concerning the Strawsonian "reversal". Such a thesis – that people are responsible because we hold them responsible – invites some immediate philosophical objections, exactly parallel to standard objections to (unsophisticated versions of) divine command theory. If something is right because God commands it, doesn't this imply, absurdly, that if God commanded murder, then murder would be right? Similarly, if people are morally responsible because we hold them responsible, doesn't this imply, just as absurdly, that if we held young children (or the severely mentally ill) responsible, they would be responsible?[3] (I return to this point below.) At any rate, this is not what Watson says, despite this being what one would *expect* Watson to say, given the parallelism at issue in standard "Euthyphro-style" contrasts concerning "explanatory priority". (Is something pious because loved by the gods, or loved by the gods because pious?) Instead, Watson says:

> The explanatory priority is the other way around: It is not that we hold people responsible because they *are* responsible; rather, the idea (*our* idea) that we are responsible is to be understood by the practice, which itself is not a matter of holding some propositions to be true, but of expressing our concerns and demands about our treatment of one another.

---

[2] Watson 1987, reprinted in Watson 2004: 221 – 222.

[3] This point has been made before. See, e.g. Fischer and Ravizza 1993: 18 – 19, Fischer 1994: 212-213, Ekstrom 2000: 148, and Nelkin 2011: 28. Note: in this paper, "mental illness" shall refer stipulatively to whatever sort of mental illness undermines moral responsibility.

And this formulation, it seems, is less clear than the one provided above. Here Watson sets up a contrast. The contrast is between partisans of the "traditional" order of explanation, such as libertarians, who endorse (1), and Strawsonians, who deny (1) and instead accept (2):

(1) We hold people responsible because they are responsible. AND

(2) The idea that we are responsible is to be understood by the practice of expressing our concerns and demands about our treatment of one another.

But now we have a set of puzzles (in a pattern that will repeat itself below). First, it isn't clear that the libertarian endorses (1). Second, it isn't clear that (1) and (2) are incompatible. And third, it isn't clear that (2) is something the libertarian would have to deny. All three results, however, are intended by Watson. In short, then, it is unclear how this passage can do the work Watson aims for it to do.

To begin, read strictly (I consider an alternative interpretation in a moment), (1) alleges that we hold people responsible because they are, in fact, responsible. On the surface, however, this is puzzling; what was (seemingly) at issue was the *propriety* of the reactive attitudes, not why we do in fact engage in them. That is, the libertarian is someone that (in some sense, or at least allegedly) invokes an "independent notion of moral responsibility" in order to explain the "propriety" of our "holding people responsible". (1), however, is a thesis concerning the explanation of why we in fact hold people responsible. Accordingly, it isn't clear what (1), as stated, has to do with what comes previously; seemingly, we have simply shifted to a different question or different topic: why we hold people responsible. At any rate, plausibly, the libertarian isn't committed to the thesis that we do in fact hold people responsible because they are responsible; rather, if anything, the libertarian is committed to the thesis that what makes it *appropriate* (all else equal) to hold people responsible is the fact that they are responsible. It is no part of libertarianism *per se* that it is people's being responsible itself that explains (presumably by causing) our holding them responsible. So, again, it isn't clear how (1) is something libertarians (or the relevant consequentialists for that matter) do or must endorse.

Further, it isn't clear how (1) and (2), as stated, are meant to be incompatible. (1) appears to be a psychological thesis concerning what explains why we hold people

responsible; (2) is a thesis concerning how "the idea that we are responsible" is to be understood. Certainly these claims do not seem to be anything like "p because q" and "q because p", which are claims we immediately recognize to be in tension (given the asymmetry of explanation). This problem would appear to persist even if we replace (1) with (1*), in accordance with the worries of the previous paragraph:

(1*) It is appropriate to hold people responsible because they are responsible.

(1*) is a claim concerning what makes it appropriate to hold people responsible. (2) is a claim concerning how "the idea that we are responsible" is to be understood. There is no obvious incompatibility between (1) or (1*) and (2).

But the chief problem for this passage is (2) itself, which is meant to be the new, Strawsonian thesis, incompatible with the libertarian (and consequentialist) approach to moral responsibility. But it is unclear how (2) could plausibly be taken to be inconsistent with libertarianism. (2) speaks of "the idea" that we are responsible. And it speaks of how this "idea" is to be understood. The claim, then, would appear to be "epistemic" in character. And it claims that this idea is to be understood "by" (which I take to mean "in the light of") "the practice" – in particular, the practice of expressing our concerns about how we are treated (which includes, presumably, our practices of blame and resentment). However, this thesis seems plainly consistent with libertarianism – and indeed, so weak as to be almost trivial. What, exactly, in libertarianism is inconsistent with the claim that the idea that we are responsible is to be understood in light of our practices of holding each other responsible? If you want to understand "the idea that we are responsible" (perhaps, what it means for someone to be responsible), then that is to be done in light of our practices; you'll have to look at our practices of blame and resentment in order to understand the idea that we are (what it is for us to be) responsible. Nothing in this thought, as far as I can see, is in in any way inconsistent with the thesis that such practices are appropriate only if we have a kind of freedom incompatible with determinism.

There is one further statement Watson makes that is worth considering, given how often it (or a similar such statement) appears in the literature. Earlier in his paper, Watson writes that

> Strawson's radical claim is that these "reactive attitudes" (as he calls them) are *constitutive* of moral responsibility; to regard oneself or another as responsible just is the proneness to react to them in these kinds of ways under certain conditions.[4] (2004: 220)

But what is it, exactly, to say that the reactive attitudes are "constitutive" of "moral responsibility"? It isn't easy to say. Is Watson here speaking of "the idea" of moral responsibility – that is, is the claim that *what it is* to be morally responsible is (say) to be a target (or perhaps an "appropriate" target?) of the reactive attitudes? Or is the claim instead a claim about what *makes* someone responsible – namely, that being subject to the reactive attitudes *makes* one responsible (deserving of such attitudes)? Or is it something else? I don't exactly know. "The reactive attitudes are constitutive of moral responsibility" is a phrase that doesn't wear its meaning on its sleeve.

Further, the remark that follows seems to be of little help. Here we have two claims:

(a) The reactive attitudes are constitutive of moral responsibility. AND (taking some liberties)
(b) One's belief that someone is responsible just is one's tendency to react to that person in certain ways under certain conditions.

Claim (b) simply *identifies* one's belief with a tendency. Consider first, however, the weaker claim, not that one's belief *is* the tendency, but that one's belief that someone is responsible *essentially manifests itself* in a tendency to react to that person in certain ways:

(b*) Necessarily, a belief that someone is responsible implies a tendency to react to that person in certain ways under certain conditions.

Note: there would not seem to be anything "radical" about a claim such as (b*), and there doesn't seem to be anything in this idea at all inconsistent with libertarianism. The

[4] Various subsequent authors have written that, on the Strawsonian view, the reactive attitudes are "constitutive" of moral responsibility (e.g. Russell 1992: 288, Haji and Cuypers 2008: 186, Campbell 2011: 8, Echeñique 2012: 22), or that moral responsibility is "constituted by" the reactive attitudes (e.g. Double 1996: 64, Haji 2002: 204, Kane 2005: 108, McKenna 2008: 202).

libertarian can plainly maintain that, necessarily, you (genuinely) believe that someone is responsible only if you are disposed to react that person in various ways in various conditions. It would thus also seem clear that libertarian can accept the *stronger* claim at issue in (b), the claim that one's belief simply *is* the relevant tendency. (Libertarianism would not seem to be committed to a particular non-dispositional analysis of belief.) In sum, it is puzzling how (b) or (b*) could plausibly be taken to illuminate the meaning of (a), and it is puzzling how (b) or (b*) could be meant to be "radical".

Kane

As I see it, I am not the only one unsure of the implications of saying that the reactive attitudes are "constitutive" of moral responsibility, or that moral responsibility is "constituted" by such attitudes. For instance, Robert Kane, in his *A Contemporary Introduction to Free Will*, writes (echoing Watson):

> Now many people have recognized these connections between free will and responsibility on the one hand, and reactive attitudes—such as resentment, blame, admiration, and gratitude—on the other. But what is unique about Strawson's theory is his belief that responsibility is *constituted* by our adopting such reactive attitudes toward one another. (108)

Consider the claim that responsibility is *constituted* by our adopting reactive attitudes toward one another. As before, what this means isn't perfectly clear. But, insofar as it has a natural meaning, it seems to me to be this: the fact that we have these attitudes towards one another *makes* us morally responsible – that is, *makes* us deserving of such attitudes, or "appropriate targets" of such attitudes. If you say that something's being red is constituted by its disposition to cause certain sensations in certain subjects, then you're saying, I think, that the fact that something causes those sensations *makes* it red. The very fact that we have these attitudes towards one another makes us morally responsible – that is, makes us fit subjects (and deserving) of such attitudes.[5]

[5] Cf. Pereboom 2013: 615: "In [Strawson's] view, … the fact that agents are typically resented for certain kinds of immoral actions is what constitutes their being blameworthy for performing them."

Now, *this* would certainly be a radical, unique claim, if this were the claim. Is it? Kane seems unsure. For he immediately adds:

> What justifies us in holding people responsible is that they are part of a *practice* or *form of life* in which it is appropriate to have such reactive attitudes toward one another.

Note: Kane does not write (as would seem expected, given the line prior) that what justifies us in holding people responsible is that they are part of form of life in which we *do* have reactive attitudes towards one another, but that they are part of a form of life in which this is *appropriate*. But then in what sense is responsibility "constituted" by our adopting reactive attitudes towards one another? Wouldn't it instead seem to be "constituted" by whatever makes such attitudes *appropriate*? At any rate, in this latter claim, not only do we not have something radical, we seem to have something that no one could reasonably dispute. What justifies us in blaming and resenting people? The fact that those people are involved in a certain form of life. What sort of form of life? The sort on which it is appropriate that they be blamed and resented. So: what justifies us in holding people responsible – blaming and resenting them – is the fact that it is appropriate that they be blamed and resented. What is "unique" to Strawson's theory turns out to be a borderline triviality.

Watson 2014 (and Brink and Nelkin 2013)

In more recent work, however, Watson returns to the themes regarding the "order of explanation" he developed in his 1987. In his new (2014) essay, Watson begins by arguing that, though Strawson's name is now continuously invoked in discussions of moral responsibility, "very few philosophers have taken a truly Strawsonian turn." And that turn, Watson says, is anti-libertarian. Watson explains:

And Haji 2002: 204: "Responsibility, then, is nothing more than – it is constituted by – our adopting these attitudes toward one another." And Double 2002: 519: "For Strawson, …"being" morally responsible just is belonging to a web of social interaction in which persons *do* hold each other responsible…". And Sher 2006: 81: "[Strawson argued] that to be responsible for one's acts is precisely to be someone whose failures to display good will are in fact prone to elicit such responses [the reactive attitudes]."

> In the past twenty years, it has become routine to characterize responsibility in terms of the propriety conditions of what Strawson calls the reactive attitudes, but this manner of speaking has not gone along with a commitment to the claims that give Strawson's essay its striking originality.
>
> To cite just one illustration, in a recent paper, David Brink and Dana Nelkin assert the following biconditional, which they dub "Strawson's thesis": "Reactive attitudes involving blame and praise are appropriate just in case the targets of these attitudes are responsible" (2013: 287). Brink and Nelkin hasten to add that, in contrast to Strawson, they endorse what they call a "realist" interpretation of the thesis, according to which being responsible is an independent property evidenced or presumed by the reactive attitudes. Strawson himself, they rightly say, understands being responsible in a "response-dependent" way; to be responsible just is to be a (possible) fit target of that sort of attitude. (16)

Watson adds:

> To be sure, the further idea that blame and praise themselves have to be understood in terms of the reactive attitudes is a controversial and distinctive part of Strawson's picture, and perhaps even original to him. But unless this idea is coupled with something like his "response-dependence" thesis, we have nothing close to a Strawsonian understanding of responsibility.
>
> It is worth noting that if seeing the reactive attitudes as conceptually central to responsibility were sufficient for being Strawsonian, then even libertarians might raise that banner.[6] This gesture may have some rhetorical point, but we should resist it on account of its theoretical superficiality. (16)

Watson here *seems* to be intimating that, once we endorse the "response-dependent" conception of responsibility at issue, then we have (at least something close to) a Strawsonian conception of responsibility – and *that* conception of responsibility is one the libertarian cannot endorse. Libertarians, of course, can (and perhaps many do) endorse the

[6] Here Watson cites Christopher Franklin's 2010 dissertation, *Strawsonian Libertarianism: A Theory of Free Will and Moral Responsibility*.

mere biconditional itself. However, libertarians cannot be on the right side of the divide regarding how it ought to be *interpreted*.

But recall Watson's discussion of the biconditional at issue. Here Watson and Brink and Nelkin set up a contrast. The contrast is between the "realist" interpretation of the biconditional, and (what they call) the "Strawsonian," "response-dependent" interpretation of the biconditional. Simplifying only to the case of blame, the biconditional becomes:

(T) Blaming a given agent is appropriate if and only if that agent is blameworthy.

Now, what is the "realist" interpretation of the biconditional? It is the interpretation on which blameworthiness is *presumed* by appropriate blame; more particularly, it would seem, it is the interpretation on which the explanatory direction moves from right to left, viz.:

(R) Blaming a given agent is appropriate *because* that agent is blameworthy.

That is, in the "order of explanation", "first" one has the target's blameworthiness, and it is *in virtue of* this blameworthiness that blame is appropriate. Now, what is the "response-dependent," "Strawsonian" interpretation of the biconditional? Presumably, it is the one on which the explanatory direction moves the other way, viz.:

(S) A given agent is blameworthy *because* blaming that agent is appropriate.

After all, consider Watson's own gloss on the thesis: "to be responsible [blameworthy] just is to be a (possible) fit target of that sort of attitude [blame]." If we read the "just is" locution to imply the relevant kind of priority (as would seem natural), then we have the following: One is blameworthy *because* one is an appropriate target of blame. That is, we have (S).[7]

But now the point. Though Watson can seem to be intimating that (S) is incompatible with libertarianism, (S) is plainly *consistent* with libertarianism. The libertarian can maintain that someone is blameworthy because it is appropriate that she be blamed, *and* that no one is appropriately blamed if determinism is true. There is not the slightest hint of

[7] For discussion on whether to endorse (R), (S), neither, or, as he would have it, both, see McKenna 2012: Ch. 2.

a contradiction between these two positions. (S) itself tells us nothing about the *conditions* of appropriate blame, and so *in itself* has no anti-libertarian implications. If taking a truly Strawsonian turn is to endorse something incompatible with libertarianism, as Watson maintains, then we will have to turn considerably farther than merely moving from (T) or (R) to (S).

Essentially the same complaints might be brought to bear on Brink and Nelkin's discussion of these issues. As Watson indicates, Brink and Nelkin note that "Strawson's thesis can be interpreted in two very different ways, depending on which half of the biconditional has explanatory priority"; Brink and Nelkin further say that they endorse the "realist" interpretation of the biconditional (corresponding to (R) above). Here they comment:

> A response-independent ["realist"] conception of responsibility is hostage to traditional worries about freedom of the will. The problem of free will is the problem of reconciling responsibility with determinism, because responsibility may seem to presuppose freedom of the will, and freedom of the will may seem incompatible with determinism. (4)

To be sure, though Brink and Nelkin concede that their realist interpretation is "hostage" to these worries, ultimately they contend that responsibility can indeed be reconciled with determinism. But the implication is clear: whereas the "realist" interpretation of the biconditional makes responsibility "hostage" to worries about determinism, the "response-dependent" conception (captured by (S)) *does not*. Once we "reverse" the "order of explanation" from (R) to (S), worries about responsibility and determinism cannot so much as arise. But this is false. Moving from (R) to (S) leaves any such worries just as they were.

As far as I can see, then, whether to endorse (what Watson and Brink and Nelson call) the "realist" or instead the "response-dependent" interpretation of the given biconditional is simply a red-herring vis-à-vis the compatibility debate.[8] Watson, it seems, is

[8] I should note that, whereas Brink and Nelkin (and Watson) call the relevant reading of the biconditional a "response-dependent" conception of responsibility, it is not clear to me that the given conception is genuinely "response-dependent". Of course, the terminology of "response-dependence" is contested (for some of the various complexities, see, e.g. Wedgwood 1997 and López de Sa 2013), but at least on some conceptions of "response-dependence" (e.g. Johnston's 1989

sensitive to the fact that (S) alone will not deliver us a compatibilist result, for he immediately adds:

> Despite the current fashion of framing questions of responsibility in terms of the reactive attitudes, then, very few philosophers have taken a truly Strawsonian turn. Let me now try to be more constructive by setting out what I regard as the elements of his project and then taking up a few of the many critical issues it raises. (16)

The suggestion, perhaps, is that Watson will *now* tell us how Strawson's project is anti-libertarian. Ultimately, however, as I hope to show, even here we encounter a great deal that the libertarian can accept – and arguably far less that she cannot.

In the following section, "The Distinctive Contribution of Strawson's Essay," Watson begins:

> What is fresh in "Freedom and Resentment", as I read it, are two related ideas: that our sense of ourselves and one another as morally responsible agents and (accordingly) as morally responsible *to one another* is integral to ("given with") human sociality itself, and that attempts to ground "responsibility practices" in some reality external to human nature are misguided. (17)

Here we have two claims:

> Our sense of ourselves as morally responsible agents is given with human sociality.

---

conception, which introduced the phrase), the relevant proposal does not fit the mold. On these views, the response-dependent conception of responsibility is not one according to which one is responsible because *appropriately* held responsible, but instead according to which (roughly) one is responsible because *disposed to be held responsible*. That is, the concept or property of blameworthiness is the concept or property of being disposed to cause feelings of blame/resentment in suitably situated subjects. So long as we grant, as we should, that even on determinism, people's actions may still be liable to cause feelings of resentment, then such a conception of "blameworthiness" of course secures compatibilism. (Cf. the discussion of Wallace's thesis (D) below.) As I see it, when Brink and Nelkin are imagining that their "response-dependent" conception of responsibility does not give rise to worries about determinism, it is because they are thinking of the conception in this more robust sense – that is, as something stronger than a mere "reversal" of (R) to (S). The shift between (S) and this more robust thesis, however, is an enormously important one.

and

> Attempts to ground "responsibility practices" in some reality external to human nature are misguided.

Consider the latter claim first. Does the libertarian attempt to "ground responsibility practices" in a "reality external to human nature"? I do not see how – at any rate, certainly the libertarian cannot be expected to agree that he or she does. If the question is what grounds the fairness (or appropriateness) of holding a particular agent responsible for a particular action, then the pertinent claim of the libertarian is that this fairness is grounded at least in part in the fact that the given agent could have done otherwise than what he did.[9] She adds that this kind of freedom is incompatible with determinism. Is the fact that we have this kind of freedom "external" to human nature? Why should it be? Why can't the libertarian maintain that our having the ability to do otherwise is, indeed, part of "human nature"? Indeed, I can imagine libertarians *arguing* that such freedom is part of human nature.

Consider now the former claim, viz. that "Our sense of ourselves as morally responsible agents is given with human sociality." What is "human sociality"? Watson explains:

> Strawson identifies two components of human sociality as crucial here. First, we care deeply (and "for its own sake") about how people regard one another. Second, this concern manifests itself in a demand or expectation to be treated with regard and good will. Following Strawson, let's call these the *basic concern* and the *basic demand* respectively. (17)

Given this understanding of "human sociality," the claim becomes:

---

[9] Here and in what follows I consider only the perspective of (classical) libertarians who believe that the ability to do otherwise is necessary for moral responsibility. So-called "source incompatibilists" (for more on which see, e.g., Timpe 2012) are free to substitute their sourcehood condition for the alternative possibilities/ability to do otherwise condition in what follows.

> Our sense of ourselves as morally responsible agents is given with the fact that we care deeply about whether we are treated with good will, and therefore demand to be treated with good will.

It should be clear that the libertarian can *accept* this claim. That is, the libertarian can accept the claim that, given that we do in fact care about how we are treated, and do in fact demand to be treated in certain ways, it follows that we will have a "sense of ourselves" as morally responsible agents. (Whether it follows from these claims that we *are* responsible agents is another matter. I consider this alternative reading in a moment.) Watson goes on:

> To be a responsible agent is to be someone whom it makes sense to subject to such a demand. (17)

This is, crucially, also a claim the libertarian can accept. Here the libertarian would simply contend that, in the end, it does not make sense to demand (or normatively expect) someone not to have done what she could not have refrained from doing, and that this kind of freedom is incompatible with determinism. More generally, our demands and expectations only make sense against the backdrop of the supposition that those subject to them possess the freedom to do otherwise. Of course, this contention is *controversial* – but the important point is that one cannot put the given claim forward as something to be *contrasted* with libertarianism. Indeed, I can imagine certain libertarians *arguing* that to be responsible is to be someone whom it makes sense to subject to such demands.

Watson continues:

> *Inter alia*, that demand displays itself in various practical and sentimental reactions to the attitudes people take to one another (to their "quality of will"). (17)

This much is uncontroversial and certainly no threat to the libertarian. He continues:

> Thus our social sentimental nature grounds the distinctive reasons that structure our personal relations, but that nature is itself rationally brute, since it provides the framework for rational scrutiny. Accordingly, Strawson argues, the traditional

> philosophical discussion of this topic is marred by attempts to "validate [or invalidate] our general disposition to moral response and moral judgment", attempts that in one way or another "overintellectualize" the phenomena. (17)

Thus concludes the given section. If we are to find something anti-libertarian about "the distinctive contribution of Strawson's essay," as we are meant to, then it will have to be here. And, if we search for it, an anti-libertarian reading of this passage is not hard to hard to find. Consider the claim – implicitly targeting the libertarian – that we should refrain from attempts to "validate" our general disposition to blame. This, it may seem, is a command the libertarian cannot obey, as a matter of the logic of her position. The pertinent claim of the libertarian is that the fairness of our blaming someone is (at least in part) grounded in the fact that this person could have done otherwise – that is, that the agent met an "avoidability" condition on being responsible. In this sense, and, arguably, in this sense *only*, the libertarian seeks to "validate" our disposition to moral response and judgment: for the libertarian, the fairness of holding an agent responsible is (at least in part) grounded in ("validated" by) the fact that the agent meets certain objective conditions. If this passage is to have an anti-libertarian upshot, then, it will have to be because the Strawsonian denies precisely this thesis: that the fairness of holding an agent responsible is grounded in facts – and here it would seem to be *any* facts – about the agent at all. Any such claim amounts to a misguided attempt to "validate" the relevant responses. Such a position, of course, is thoroughly anti-libertarian. But it is also, I believe, thoroughly implausible.

My claim here is not that Watson (or the Strawsonians he means to represent) would be happy to concede that the fairness of holding an agent responsible is entirely ungrounded in this sense – that is, that there simply *are no* objective conditions one must meet in order fairly to be blamed. Indeed, in various other parts of his essay, Watson freely speaks precisely of these conditions. My claim, instead, is that Watson here faces a dilemma – a dilemma that lies at the heart of the present paper. In order to make the Strawsonian conception articulated above plausibly anti-libertarian, one must make it implausible on other grounds. That is, one could affirm that, even for the Strawsonian, there are such objective conditions; merely providing objective conditions an agent has to meet in order fairly to be blamed does not constitute a problematic attempt to "validate" our given responses. In that case, however, the conception is not anti-libertarian – or anyway not yet.

If one allows such conditions at all, then there is nothing *yet* stopping them from being libertarian. On the other hand, one could affirm the radical view that there simply *are no* conditions a given agent must meet in order fairly to be blamed. And, of course, if there are no such conditions, then there are *ipso facto* no *libertarian* such conditions. The problem with this conception, however, is simple. If the fairness of blame isn't "validated" by something about the relevant agent, and yet blaming her is nevertheless fair, then we might wonder: wouldn't it follow that blaming *just anyone* could be (or could have been) fair? For instance, what resources could the Strawsonian bring to bear to argue that, if we had been inclined to blame (and did blame) the mentally ill and small children, they nevertheless would not be fairly blamed? They could not, of course, appeal to the fact that such persons fail to meet a given requirement on being fairly blamed – for to provide any such requirement is thereby to (attempt to) "validate" our blaming responses. The Strawsonian seems stuck: one can rule out libertarian conditions on responsibility (by ruling out any such conditions whatsoever), but only at the expense of accepting the given result about the mentally ill and small children.

Watson goes on to provide a summary statement of "the elements of the picture" at issue, namely

> that (1) responsibility is "given with" our social nature, (2) which is constituted by the basic concern, (3) for which it is wrong-headed to seek a rational grounding, and that (4) there is no further fact about responsible action and agency beyond the realized capacity for inter-personal relations to which our responsibility practices are answerable; in this sense, to be a responsible agent is to be appropriately subject to the basic demand. (18)

I consider these claims in turn. First, note: here Watson has moved from saying that our *sense of ourselves* as responsible is "given with" our social nature to the claim that *responsibility* is "given with" this nature. But there is a crucial difference between our *sense of ourselves* as responsible and our *being* responsible. On this new construal, the claim encapsulated by (1) and (2) would seem to be:

> Given that we care deeply about how we are treated (the "basic concern"), it follows

that we are morally responsible.

But this thesis, even if denied by the libertarian, anyway seems implausible. How should it follow from the fact that we *care* (however deeply) about whether we are shown good or ill will that we are responsible for which sort of will we show? That we care how we are treated is one thing; whether anyone is *responsible* for treating us a given way seems to be another.

Now consider the claim (at issue in (3)) that there is something for which it is "wrong-headed" to seek to give a "rational grounding". As we just saw, the pertinent claim of the libertarian is that the fairness of blame is grounded in facts about the relevant agent. If (3) is to be anti-libertarian, then, this is the claim that would have to be denied. As we saw, the Strawsonian may certainly deny this claim – but then she encounters the given difficulties about the mentally ill and small children.

Consider finally the (slightly simplified) claim at issue in (4), viz.:

> There is no further fact about responsible agency beyond the capacity for interpersonal relations.

I interpret this claim to mean that

> There is no further condition one must meet in order to be morally responsible beyond having the capacity for interpersonal relations.

But the capacity for interpersonal relations of *which sort*? Watson clarifies: the capacities needed to be "appropriately subject to the basic demand." As we saw, however, once clarified in this sense, the libertarian can accept the claim that having the capacity for interpersonal relations is sufficient for responsible agency – for having this capacity implies being appropriately subject to the basic demand, and being appropriately subject to the basic demand implies being free to do otherwise. Once the "capacity for interpersonal relations" is understood in the given way, the libertarian will maintain that this capacity already encodes her conditions on morally responsible agency.

In the remainder of his essay, Watson goes on further to develop what he calls the "normative standpoint argument", the basic thrust of which is that our practices of holding

responsible are normatively basic, in the sense that the fairness of these practices is properly immune to criticisms appealing to standards external to those practices themselves. Watson, however, implicitly concedes that the libertarian can *accept* this point. As he says,

> Some will say that Strawson misinterprets the commitments implicit in the reactive attitudes, that, contrary to his arguments, we find on reflection that they implicate propositions that are false or at least very doubtful if determinism is true (or even if it's not true). The framework contains the seeds of its own refutation. (24)

Some will say – and the libertarian *does* say. According to our own moral practice, it is not fair that someone should be blamed for what she had no power to avoid. In that sense, according to the libertarian, if determinism is true, the framework contains the seeds of its own refutation. If the above is a fair reply to the normative standpoint argument – and Watson concedes that it is – then, once more, we do not here have anything the libertarian cannot accept. Instead, we have a disagreement concerning the standards implicit in our practices. Is it fair that anyone should be blamed for what she had no power to avoid? The libertarian says it isn't, and thus that determinism rules out responsibility. The compatibilist says that it is, or instead that determinism is no threat to avoidability. We are back to the familiar debates.[10]

The point, then, is clear. If we are to understand the way in which Strawson "reversed" the "order of explanation" regarding moral responsibility – and if we are to have a Strawsonian conception of the "facts of moral responsibility" that is plausible, and *also* plausibly anti-libertarian, then we will have to look elsewhere than that provided by Watson (and, in turn, Kane and Brink and Nelkin).

---

[10] Cf. Gideon Rosen's (2002) contribution to a book symposium on Wallace 1994, who considers libertarianism on precisely these grounds:

> Perhaps you can look Judas in the eye and hold him responsible for his betrayal, knowing full well that while he satisfies Wallace's conditions, it was settled from the beginning, as a matter of necessity, that he would act badly in the circumstances. Well then, we disagree about what is fair. Could this be bedrock? That would certainly be disappointing. (2002: 706)

Disappointing, perhaps. But there it is.

Coates and Tognazzini

In the Introduction to their *Blame* book, Coates and Tognazzini present the Strawsonian view in some detail. They begin:

> The details of Strawson's proposed alternative [to the utilitarian justification of blame and the libertarian's way of justifying blame] are controversial, but here's the basic idea. Instead of viewing blameworthiness as an independent metaphysical fact about an agent (or based on such a fact), as the libertarian does, the utilitarian is right to view it as somehow essentially tied to our blaming practices. (6)

First, however, it looks like there are some senses in which libertarians would *agree* that blameworthiness is "essentially tied to our blaming practices". For instance, arguably, blameworthiness is (for the libertarian) that thing that (at least in part) justifies or makes appropriate those practices. Isn't that a sense in which blameworthiness is essentially tied to those practices?

They go on:

> But the libertarian is right to insist that our blaming practices are more than instruments for the regulation of behavior. As Strawson puts it, "Our practices do not merely exploit our natures, they express them". And the relevant aspect of human nature is "that complicated web of attitudes and feelings which form an essential part of the moral life as we know it", namely the *reactive attitudes* of resentment, indignation, and guilt (among others). These attitudes are precisely what is left out of the utilitarian picture of blame, but according to Strawson, "it is just these attitudes themselves which fill the gap" and not some mysterious appeal to metaphysical freedom. To be morally responsible, on this account, just is to be a member of the moral community, to be someone toward whom others feel the reactive attitudes. (6)

In this final sentence, we have what seems to be a straightforward articulation of the point one *expected* Watson to make above, but which in fact he did not make: To be morally

responsible (blameworthy) just is to be someone whom others hold responsible (blame). On the most natural way of understanding the "just is" locution, we could say: What makes someone blameworthy is the fact that others blame him. But this view is subject to the serious difficulty noted above: it would seemingly imply that if we blamed the severely mentally ill, they would be blameworthy, and that seems false. Again, we have a difficulty exactly parallel to the familiar Euthyphro objection to divine command theory.

They go on to say that

> [P]erhaps the most common way of conceiving of moral responsibility these days is along broadly Strawsonian lines, emphasizing the importance and explanatory priority of our practices of blaming and holding one another responsible. These practices (together with their associated norms) are not (taken to be) constrained by any independent "moral responsibility facts" about the agent in question; rather they are what partly determine which facts about an agent even *count* as the moral responsibility facts in the first place. (6)

Here again we are meant to have a contrast. The contrast would seem to be between

> (3) Our blaming practices are constrained by independent "moral responsibility facts" about agents. AND

> (4) Our blaming practices partly determine which facts about an agent count as the moral responsibility facts for that agent.

However, once more, it is not clear that the libertarian does or must accept (3), it is not clear that (3) and (4) are incompatible, and it is not clear that the libertarian denies (4), whereas all three results are intended by Coates and Tognazzini.

Consider (3). What (3) means is not entirely clear. What are "moral responsibility facts" about agents? I am not claiming that there is no good reading of (3) on which the libertarian endorses (3); my claim only is that the libertarian certainly could not be expected immediately to recognize (3) as a claim that she does or must endorse. Further, is (3) indeed

incompatible with (4)? Again, I don't immediately see how; part of the difficulty is that what (3) and (4) mean is not entirely obvious.

More importantly, (4) is meant to articulate the new "Strawsonian" order of explanation concerning moral responsibility, incompatible with the traditional understanding paradigmatically endorsed by the libertarian. However, must a libertarian deny (4)? I don't see why. I don't see why a libertarian can't accept the idea that our blaming practices partly determine which facts about an agent count as the moral responsibility facts for that agent. (This is in part because I have difficulties understanding what this claim means.) There is, in any case, at least one interpretation of (4) that the libertarian could endorse. Suppose we ask: what are the "moral responsibility" facts about Bob? Is he a pretty blameworthy guy, or instead pretty praiseworthy, or what? (Is that what the "moral responsibility facts" are for him?) Well, part of what determines whether Bob is a blameworthy guy is whether he's worthy of (deserving of) blame. What's blame? Well, blame is something we direct towards other people – if you want to know what blame is, you'll have to look at our blaming behavior to find out. Is this a sense in which, for the libertarian, our blaming practices determine which facts about an agent count as the "moral responsibility facts" for that agent? Our practices determine which thing is in question: the desert of *that kind* of response (blame). (The libertarian simply adds: *that kind* of response is appropriate only if agents possess a kind of freedom incompatible with determinism.) Does this suffice for an acceptance of (4)? Perhaps Coates and Tognazzini would say that it doesn't. If so, then I am simply unsure what (4) implies.

Finally, in a footnote following the above passage, they add:

> This is not to say that our practices are not constrained by any facts about the agent whatsoever. The reactive attitudes may still be inappropriate if the agent lacks certain crucial capacities. (6)

But now we have another puzzle – a puzzle (it seems to me) not adequately addressed in Coates and Tognazzini's discussion. It certainly looked initially (in the passage noted above) like we were getting the radical claim: blameworthy if blamed (indeed, because blamed). Now we're told otherwise: you could be blamed and yet not blameworthy, if you lack a certain capacity. But isn't this precisely to admit that our practices are indeed

constrained by "moral responsibility facts" about the agent in question, viz., whether she has the relevant capacity? What seemed like was given with one hand is taken back with the other.

Tognazzini

As I see it, this same problem arises once more in Tognazzini's (individual) 2013 discussion of the Strawsonian "reversal". Tognazzini writes that Strawson

> proposed an unfamiliar but incredibly intriguing conceptual reversal: rather than viewing our practices of holding each other morally responsible as answerable to independent facts of moral responsibility, we should view the facts of moral responsibility as answerable to – or, at least, as partly determined by – those practices. (1300)

This statement is, I think, relatively straightforward, but for one complication: the "partly" qualifier on "determined". But for that qualifier, we could understand the "reversal" on analogy with divine command theory:

> Rather than viewing God's commands as answerable to independent facts about what is right and wrong, we should view facts about what is right and wrong as determined by God's commands.

After all, the divine command theorist does not contend that God's commands only *partly* determine facts about what is right and wrong, but that God's commands *wholly* determine what is right and wrong. Indeed, given the first clause of the statement, it is simply unclear what it could mean to say that God's commands may only "partly" determine what is right and wrong. At any rate, here is one way in which one could *not* satisfactorily attempt to make use of such a "partly" qualifier on "determined". Suppose someone says:

> Rather than viewing God's commands as answerable to independent facts about what is right and wrong, we should view facts about what is right and wrong as partly determined by God's commands.

And suppose we said the following in reply:

> Look, if God's commands are in no way "answerable to" any facts about what is ("already") right or wrong, then couldn't God have simply commanded *anything* to be right or wrong, and then it would be?

And suppose we received this reply:

> No. This is because I said that God's commands only *partly* determine what is right or wrong. I'm not saying that God could command *just anything* to be right or wrong and it would be right or wrong. There are constraints on what God may permissibly command to be right or wrong.

If this is the reply we received, then I think we would suspect this person of speaking out of both sides of his mouth. After all, he *began* by saying that God's commands are not "answerable to" independent facts about what is right or wrong. Now, however, when challenged with the given (seemingly absurd) implication, he invokes the claim that there *are* normative constraints on what God may command – an expedient his "partly" qualifier is (somehow) meant to allow.

As I see it, precisely this sort of problem plagues Tognazzini's (and Coates and Tognazzini's) development of the Strawsonian "reversal": the "partly" qualifier is somehow meant to block what *seems* to be the absurd implication of Strawsonian "reversal", but it is unclear how it can successfully do so. Tognazzini nowhere addresses this challenge head on, but (arguably) its presence is in the background of his development of the view. The result, I believe, is a theory that is simply not very easy to understand. For instance, Tognazzini writes:

> According to the Strawsonian picture … there are no independent facts of moral responsibility. Now, this is not to say that there are no facts *at all* about moral responsibility – just none that is independent of our practices of holding each other responsible. Introducing his conceptual reversal, Strawson takes our practices of blame and holding responsible to be what fixes the facts of moral responsibility. (1301)

On the one hand, it can seem that the proposal is that our practices of holding responsible "fix" the facts of moral responsibility just as, on divine command theory, God's commands "fix" the facts of what is right and wrong. Accordingly, then, it would seem to follow that if we blamed (or were disposed to blame) the severely mentally ill, for instance, they indeed would be blameworthy, just as it seems to follow (from divine command theory) that if God commanded murder, murder would be right. On the other hand, Tognazzini writes that there indeed are "facts about moral responsibility" – just no *independent* such facts. But I am unsure how to interpret this claim; perhaps, however, the idea is simply that those facts will be *constituted* by our practices of blame and holding responsible, in the same way that (on divine command theory) God's commands *constitute* what is right. Tognazzini later writes that the relevant practices

> will not be completely *unconstrained* by relevant facts about the agent and her wrongdoing – excuses and justifications are still important, as is control, perhaps – but the crucial point is that they are not constrained by some prior and independent facts of moral responsibility or blameworthiness. On the Strawsonian view, there just are no such prior "responsibility facts", but that doesn't mean there are no prior facts at all. And it certainly doesn't mean that the practices of the moral community can be justified *completely independent* of what's true about the agent, intrinsically considered. (1311)

I am not sure what Tognazzini means to suggest when he writes that, for the Strawsonian, there are no prior "responsibility facts" (presumably, in this context, objective conditions one must meet in order to be morally responsible to which the appropriateness of our practice of "holding responsible" is "answerable"), but that there are indeed "prior facts" of

some sort. Do these latter "prior facts" imply that if we were to blame (or be disposed to blame) the mentally ill or small children, they nevertheless would not be blameworthy? If so, it is difficult to see the sense in which our practices "fix the facts" of moral responsibility. As far as I can see, Coates and Tognazzini and Tognazzini have tried to articulate a Strawsonian "middle ground" on which our blaming practices *determine* who is blameworthy, but nevertheless are "answerable to" (for instance) facts concerning what sort of control we have over our behaviour. If we lack this sort of control, then we aren't blameworthy – apparently even if we are in a moral community undisposed to regard this sort of lack of control as an excuse. But then it is simply unclear how such dispositions to blame *determine* who is blameworthy. The "middle ground", it would seem, is essentially unstable.

Wallace

There is one final source any adequate discussion of these issues should investigate: R. J. Wallace's highly influential development of "Strawsonian" themes in his 1994 *Responsibility and the Moral Sentiments*. The "reversal", of course, is in some sense a claim (or meant to be a claim) about the *order of explanation* regarding moral responsibility – a claim about "what depends on what". In Chapter 4 of *Responsibility and the Moral Sentiments*, Wallace seeks to give a particular "interpretation" of the "facts of moral responsibility" – an interpretation on which such facts "depend" (in a certain sense) on the "practice of holding responsible". It can seem, then, that Wallace is suggesting something deeply similar (if not identical) to what others have meant by "the reversal". As we will see, however, Wallace's points do not easily map onto any sort of "reversal" thesis – and, importantly, Wallace never identifies libertarians (or incompatibilists more generally) as taking the "wrong" side of the divide he articulates. His discussion thus has a rather different flavour than those considered above.

To explain. Suppose we want to know the conditions under which someone is morally responsible for what he or she does. How should the facts about these conditions be understood? On what Wallace calls the "metaphysical interpretation" of these facts, we "postulate facts about responsibility that are completely prior to and independent of our practice of holding people responsible." That is,

> we would suppose that there is a fact of the matter about responsibility "in itself", a fact about what it is to be genuinely or really responsible, and that this fact is prior to and independent of our practice of treating people as morally responsible agents. That practice would then be in good order to the extent that it succeeds in tracking or meshing with the prior and independent facts about moral responsibility. (87 – 88)

However, Wallace thinks that such an approach is "most unpromising," maintaining that he "cannot see how to make sense of the idea of a prior and independent realm of moral responsibility facts" – a view that has such facts "inhering in the fabric of the world completely independently of our activities and interests." (88) Thus, to avoid the "metaphysical interpretation", Wallace says, we must "interpret the relevant facts as somehow dependent on our practices of holding responsible." (89) After rejecting one proposal about how we might do so, he offers what he calls his "normative interpretation" of the relevant facts:

> (N) *S* is morally responsible (for an action *x*) if and only if it would be appropriate [where Wallace goes on to say that this means "fair"] to hold *S* morally responsible (for action *x*). (91)

Wallace sees (N) as providing a "schema for understanding how there could be facts about moral responsibility that depend in some way on our practice of holding people responsible," and that this is

> achieved without postulating a prior and independent realm of moral responsibility facts. Instead, the facts by reference to which the debate is to be decided are specified in terms of our practice of holding people responsible: they are facts about whether it would be appropriate [fair] to adopt toward people the stance of holding them responsible, if determinism is true. (91 – 92)

This is, in effect, Wallace's "Methodological Interlude". According to Wallace, the debate becomes the following: the incompatibilist alleges that there is something inherent in the

moral standards that we already accept that implies that, if determinism is true, it is never fair that anyone should be blamed. The incompatibilist seeks to "generalize" (employ what Wallace calls the "generalization strategy") from principles undergirding certain excuses and exemptions we already recognize to a principle of alternative possibilities – alternative possibilities we would, according to the incompatibilist, universally lack under determinism. In Chapters 5 and 6, however, Wallace contends that such a strategy fails: our standards of fairness in holding people responsible do not in fact implicate any principle of alternative possibilities. Compatibilism is thereby vindicated.

Now, it is certainly beyond the scope of this paper to *evaluate* Wallace's response to the "generalization strategy" at issue. The important point, for our purposes, is the following. Wallace's "normative interpretation" of the facts of responsibility is not meant to *settle* the debate between the compatibilist and incompatibilist. Rather, it simply *sets the stage* for that debate. There is no suggestion whatever that the incompatibilist cannot accept principle (N) and the "normative interpretation" at issue; indeed, there is precisely the *opposite* contention, viz., that the incompatibilist should *accept* (N), and *go on* to argue that the generalization strategy in fact succeeds. Thus, we cannot suppose that, according to Wallace, the libertarian/incompatibilist does not (or cannot) accept his favored interpretation of the "facts of moral responsibility". His discussion is thus importantly different than those considered above. In the discussions considered above, libertarians are identified as those taking the "wrong" side of the given divide. There is no indication, however, that libertarians have to be on the wrong side of Wallace's "divide" between the relevant "normative" and "metaphysical" interpretations of the "facts of responsibility." If Wallace's discussion has defects, then, they are not the defects plaguing those I have considered above.

But what of Wallace's discussion itself? As I see it, if there is a problem for Wallace, it is that it isn't clear how he *does* succeed in avoiding the problems of the "metaphysical interpretation". Recall the proposal. Wallace contends that we avoid the "metaphysical interpretation" because the relevant facts "depend on the practice, insofar as [they are being] specified essentially in terms of it". (91) But now we may ask: how does this sort of "dependence" — something's being "specified essentially in terms of" something else — solve the problem? It isn't clear. Consider our (now familiar) analogy. Suppose a particular theist — quite naturally — finds it strange that there could *just be* a realm of facts about what is right or wrong, prior to and independent of God's will. How could there just be such

facts — facts with which God would simply be *confronted*? Suppose this theist — recoiling from a picture on which God is "confronted by" the ethical facts — sought to make such facts *dependent* on God's will. And suppose such a theist thought to capture this sort dependence by putting forward the following "schema":

> (G) An action of type T is morally right (wrong) if and only if it would be appropriate ["fair"] for God to command (prohibit) actions of type T.

And now she says that the problem is solved, for "the facts about what is right and wrong are being specified essentially in terms of God's commands." Has this theist avoided commitment to any mysterious "ethical facts" that obtain prior to and independent of God's will? I don't see how. For now we're simply appealing to *different* facts: facts about when it is *fair* for God to command something. And *those* facts are just going to *be there*, in the same mysterious way as before. We haven't really gotten anywhere — or anyway, wherever we've gotten, it's still a place in which there "just are" ethical facts prior to God's will. In other words, on this picture, it would be misleading to contend that "rightness and wrongness" are being specified essentially in terms of God's commands. More fundamentally, they are being specified essentially in terms of the *fairness* of God's commands. And these things are importantly different.

Similarly, it isn't clear how Wallace has avoided the problems of the "metaphysical interpretation" merely by providing schema (N), any more than the theist avoids commitment to ethical facts prior to God's will merely by providing schema (G). What would help, of course, is if we took out the "fairness" qualifiers in (N) and (G), and just said that something is right if and only if God *does* (or perhaps is *disposed* to) command it, or that someone is morally responsible if and only if we *do* (or perhaps are *disposed* to) hold her responsible. In fact, this is precisely the conception of the way in which the facts of moral responsibility might "depend on the practice of holding responsible" that Wallace first considers and then rejects:

> (D) *S* is morally responsible (for action *x*) if and only if we are disposed, under favorable conditions, to hold *S* morally responsible (for action *x*). (89)

The problem for (D), Wallace says, is that it "does not allow us to formulate perspicuously the incompatibilist position"; the incompatibilist, he notes, does not (or need not) maintain that we would not in fact be disposed to hold people responsible, if determinism is true. (90) Rather, their contention is that this would be *incorrect* or *unfair*.

From a certain sort of perspective, however, that a conception of the facts of moral responsibility would render the incompatiblist position a nonstarter would not seem to be a *liability* of that conception, but an *advantage* of that conception. The fundamental problem for (D), arguably, is *not* that it immediately renders incompatibilism moot (even if this is indeed a problem), but that it encounters the familiar difficulties that plague the parallel account of rightness and wrongness in terms of God's commands. Once we remove the relevant "fairness" qualifier from (G), we seem to encounter the problematic implication that were God to command murder (or, given a modified analysis, disposed to command murder), murder would be right. Similarly, once we remove the given qualifier from (N) (and move to account like (D)), we encounter the problematic implication that, were we disposed to blame the mentally ill and small children, they indeed would be blameworthy. Neither such implication seems attractive.

Some suggestions

Wallace, then, does not articulate any clear "reversal" thesis, even if his discussion does ultimately encounter problems similar to those that we have considered above. If we are to understand the "reversal", then, we will, again, need to look elsewhere. Where else might we look? At this stage, I think that a certain degree of scepticism is warranted that any clear articulation of the given thesis exists to be found. That's all very well and good, you might think: no one, so far, has managed to explain the relevant view in a satisfactory way. Can I do any better? I certainly don't claim that I can. But what is needed, I think, is what is almost entirely and conspicuously *absent* in these discussions: some examples with which the Strawsonian proposal might profitably be compared. One does not see exponents of the Strawsonian thesis saying, for instance, that our practices of holding people responsible are "prior to" or "fix" or "determine" the facts about who is blameworthy (or…) in the same sense in which some other thing is "prior to" or "fixes" or "determines" something else. This is why the analogy with divine command theory is, I think, so instructive (and so

revealing). We need some *other* examples of the phenomenon the Strawsonian sees as applying to moral responsibility. And we need to see how the Strawsonian proposal is, in some way or other, parallel to *other* similar proposals in other philosophical contexts.

I offer one example: a comparison with the concept of *funniness*. The example follows a familiar philosophical dialectic – and it is, for that reason, especially instructive.

Moral responsibility and funniness

Here's a question. Could it turn out, somehow, that nothing we have ever been laughing at has ever *really* been funny? Could some discovery – empirical or otherwise – have this sort of result? It seems plausible that the answer is 'no'. For you might think, plausibly, that what is funny is simply *determined by* what we're inclined to laugh at.[11] Consequently, nothing could "show" that nothing has ever been funny without also showing that we've never been laughing – and how could something show something like that? It is, in some sense, simply *evident* that we have been laughing. And if what is funny is determined by what we're inclined to laugh at, then it is, in just the same way, evident that some of what we have been laughing at has been funny. Or look at it this way. What view of "the funny" would we have to presuppose in order to render it an intelligible possibility that nothing we've ever laughed at has been funny? Seemingly, one would have to think that there are, logically and explanatorily prior to there being any human beings at all, simply "objective standards" concerning what is funny and what isn't – written, as it were, in Plato's heaven. On such a view, there *just are* the facts about what is funny and what isn't, and the question will be whether any agents come along who either discover such facts or whose laughing practices track them. But this seems to be an odd theory of "the funny". Accordingly, it seems better to view "the funny", not as somehow "given" explanatorily prior to our practices and inclinations, but as determined or constituted *by* them. And the result is that nothing could imply that our practices *radically fail* to "match" the facts concerning what is funny. Those practices are *fixing* the facts about what is funny.

[11] Of course, sometimes one is moved to laugh (e.g., by tickling), but not by what one takes to be comically amusing. Moreover, sometimes one is genuinely comically amused by something, but does not laugh (e.g., one merely smiles, or stifles one's laughter). For simplicity, I set these issues aside. For more on the distinction between laughter and comic amusement, see Carroll 2014a.

That, anyway, seems like an understandable, attractive point of view. But there's a problem. Can't we laugh at what isn't funny? Consider racist jokes. Certainly in such contexts we often say things such as that, though someone was laughing at something, what she was laughing at simply wasn't funny. "What you're laughing at isn't funny" seems like a perfectly sensible rebuke. As Crispin Wright says in this context, "There is nothing funny about what happened at Chernobyl," and, we might add, there is nothing funny about slavery or genocide, even if certain people *think* these things are funny.[12] But now we seem stuck. For we had just said that our practices of laughter (or facts about what we're inclined to laugh at) *determine* what is funny. The uncomfortable implication would seem to be that, if we did (or were inclined to) laugh at genocide, then genocide would be funny. But that seems like the wrong thing to say. Genocide wouldn't be funny, even if we were all inclined to laugh at it. If we were all inclined to laugh at genocide, this would say more about us (and our moral cravenness) than it would about the amusing and humorous qualities of mass murder. Must we return after all, then, to the Platonic theory of "the funny"? We seem caught between two unacceptable alternatives: embrace such a theory, or embrace the result that if we were all inclined to laugh at genocide, genocide would be funny. Neither result seems attractive.

Can there be a middle ground? Can we reject any "Platonic" facts about what is funny, and say instead that what is funny is determined by what we're inclined to laugh at, but *not* in such a way that would generate the problematic conclusion (or one like it) that if we were to be inclined to laugh at genocide, genocide would be funny? This would certainly be a neat trick, if it could be done. I, for one, don't know how to do it. But it seems like something eminently worth trying to do. If this is your project: all the best to you.

Precisely this familiar sort of dialectical situation seems to me to be underlying (at least some) discussions of "Strawsonian" theories concerning the "order of explanation" regarding moral responsibility. Just as we might ask whether something could reveal that nothing we have ever laughed at has been funny, we might also ask whether something could reveal that no one we've ever blamed has been blameworthy. The incompatibilist, it would seem, says "yes" – something indeed could show that no one has ever been blameworthy, namely, the truth of determinism. (I do not believe that it is any part of incompatibilism [or any part of any theory worth believing, for that matter] that the truth of

[12] Wright 2003: 32. For more on these issues, see, e.g., Carroll 2014b, and Gaut 1998 and 2007.

determinism could be "discovered" or "revealed"; nevertheless, the incompatibilist is committed to the result that if determinism were true, no one is in fact blameworthy.) However, just as one might think that nothing could imply that nothing we've found funny has been funny without implying that we have never been laughing, so one might think that nothing could imply that no one has ever been blameworthy without implying that we have never been blaming anyone. For one might think that, just as what is funny is *determined* by what we're inclined to laugh at, so what is blameworthy is *determined by* what we're inclined to resent. To suppose otherwise would be to suppose that there "just are" facts about what it takes to be blameworthy, logically and explanatorily prior to there being anyone at all who *blames*. And some may find these sorts of "Platonic facts" about blameworthiness to be just as mysterious – and just as incredible – as "Platonic facts" regarding what *just is* funny and what isn't.[13]

But, as we saw, there's a problem. The alternate, non-Platonic theory of "funniness" would – at least initially – seem to imply that if we were inclined to laugh at genocide, genocide would be funny. Similarly, though the "Strawsonian" proposal has what may seem to be the salutary effect of effectively "insulating" blameworthiness from the possible truth of determinism (since determinism wouldn't imply that we haven't been blaming anyone), it does so at the considerable expense of seeming to imply that if we were to blame (or be inclined to blame) the mentally ill or small children, they indeed would be blameworthy. Again, this would seem to be the plain implication of the suggestion that our blaming practices fix the facts of moral responsibility. There are, then, two options for the Strawsonian reversal theorist. The first is to embrace the result that if we were to blame the mentally ill, they indeed would be blameworthy, and somehow explain why this result is not as bad as it seems. The second is to hope for the middle ground – the ground on which though our practices somehow "fix the facts" regarding who is blameworthy and who isn't, in such a way that the truth of determinism could not show that no one is blameworthy, they do *not* do so in such a way as to imply that if we were to blame the mentally ill, they would be blameworthy. The practices need to fix the facts in a *strong* enough way so as to imply that determinism is no threat, but in a *weak* enough way so as not to imply the result about the mentally ill. Can any such middle ground be made out? I don't know – anyway, I don't

[13] Wallace, in his discussion of the "metaphysical interpretation" considered above, seems to expresses precisely this sentiment.

see how to make it out. However, as I said above regarding the parallel theory of "funniness": if this is your project, then all the best to you.

Conclusion

So where are we? If I am right that Strawsonian "reversal theorists" wish to see the facts about moral responsibility as some may wish to see the facts about "funniness" (that is, as determined by our practices or inclinations in the relevant way), then we certainly cannot *yet* conclude that the given "reversal" is simply false or otherwise on the wrong track. We can only conclude that, as developed thus far, the reversal remains *problematic*. Nothing I have said so far rules out an account of the facts of responsibility that stakes out the middle ground in question. Providing such an account is, I believe, the challenge for those who wish to take up "the reversal".

Further, suppose we maintain that this ground in fact *cannot* be made out. It is crucial to note that this fact would leave *other* "Strawsonian" arguments just as they were. For instance, consider what Russell (1992: 288) and Wallace (1994: 96) call Strawson's "naturalist" argument, an argument that proceeds from the *inescapability* of our having the reactive attitudes to the conclusion that it is not the case that those attitudes would be inappropriate under determinism. Details aside, it seems clear that *this* argument (though certainly controversial[14]) does not depend on some thesis to the effect that our responsibility practices "fix the facts" of moral responsibility in the way in which our laughing practices may "fix the facts" regarding what is funny. Rather, it depends on the thesis that practices we are inescapably engaged in cannot turn out to be (universally) morally inappropriate, as the incompatiblist would seem to assume; practices that are inescapable and fundamental aspects of our form of life are *ipso facto* globally morally appropriate. To be sure, such an argument may give our practices (or the inescapability of our practices) a kind of "priority" in the explanation of our being responsible: because the practices cannot be given up, it follows that they are (globally) appropriate. In some sense, then, the very inescapability of our practices of blame grounds our being appropriate targets of those practices – that is,

[14] For criticism of this argument, see, e.g., Russell 1992, Wallace 1994: 97 – 99, and Russell 2011: 206. For a perceptive discussion, see Tognazzini 2014.

grounds our being blameworthy.[15] Because our practices do not have the sort of explanatory priority in our being responsible as envisaged in the analogy with "funniness", it would *not* follow that they cannot have *this* sort of priority. Importantly, then, not all "Strawsonian" themes are automatically compromised by the unavailability of the "middle ground" as described above. Insofar as one identifies the heart of Strawson's "project" with the argument from the inescapability of our practices, it would follow that Strawson's project can survive the failure of "the reversal".

In conclusion, it is worth recapping where we've been in this paper. As Tognazzini has written, Strawson seems to have proposed an "incredibly intriguing conceptual reversal" regarding moral responsibility. As intriguing as this suggestion seems to be, my main point in this paper has been that, so far, the proposal has not been developed or stated in a satisfactory way. The proposal is meant to be incompatible with libertarianism, but, as we saw, when investigated more carefully, the relevant statements turn out to be either too difficult to assess or perfectly compatible with the libertarian's core thesis. There *is* something here, however, or so it (often) seems to me. And I have concluded by suggesting an analogy that I believe helpfully illuminates the basic thrust of the attempted "reversal". The upshot is this. Defenders of the Strawsonian "reversal" either need to embrace the difficult consequence that if we blamed the severely mentally ill or small children, they would be blameworthy, or instead articulate precisely the way in which their proposal escapes this result, while still being a theory on which our practices "fix the facts" of moral responsibility. They need to explain precisely the way in which the "middle ground" is, indeed, stable. Again, I certainly do not say that I know how to do it. My main point in this paper, however, is that by the looks of it so far, no one else does either.

References

Brink, David and Dana Nelkin. (2013). "Fairness and the Architecture of Responsibility,"

[15] This is not to say that what grounds Jones's being blameworthy for a particular bit of wrongdoing on a particular occasion is that we are inescapably committed to blaming *him* for his wrongdoing on this particular occasion. What grounds Jones's being blameworthy for this particular bit of wrongdoing may be that he acted with ill will, fully aware of what he was doing, and so forth. Rather, it is to say the following. Why is acting in this way – with ill will, fully aware of what one is doing – sufficient for blameworthiness? Answer: because we are, in general, inescapably committed, in virtue of our very form of life, to blaming those who act in such ways. This gives us a sense in which Jones's being blameworthy on this occasion would be grounded in the inescapability of our practices of blame.

Oxford Studies in Agency and Responsibility volume 1, ed. D. Shoemaker. Oxford: Oxford University Press.

Campbell, Joseph Keim. (2011). *Free Will*. Malden, MA: Polity Press.

Carroll, Noël. (2014)(a). *A Very Short Introduction to Humour*. Oxford: Oxford University Press.

__ (2014)(b). "Ethics and Comic Amusement," *British Journal of Aesthetics* Vol. 54 Number 2, pp. 241–253.

Coates, Justin and Neal Tognazzini, eds. (2013). *Blame: Its Nature and Norms*. Oxford: Oxford University Press.

Double, Richard. (1996). *Metaphilosophy and Free Will*. Oxford: Oxford University Press.

__ (2002) "Metaethics, Metaphilosophy, and Free Will Subjectivism," in Kane, ed. 2002, *The Oxford Handbook of Free Will*, pp. 506 – 528.

Echeñique, Javier. (2012) *Aristotle's Ethics and Moral Responsibility*. Cambridge: Cambridge University Press.

Ekstrom, Laura Waddell. (2000.) *Free Will: A Philosophical Study*. Boulder: Westview Press.

Fischer, John Martin. (1994). *The Metaphysics of Free Will*. Oxford: Blackwell Publishers.

Fischer, John Martin and Mark Ravizza, eds. (1993). "Introduction," *Perspectives on Moral Responsibility*. Ithaca: Cornell University Press.

Gaut, Berys. (1998). 'Just Joking: The Ethics and Aesthetics of Humour', Philosophy and Literature 22, 51–68.

__ (2007). *Art, Emotion, and Ethics*. Oxford: Oxford University Press.

Haji, Ishtiyaque. (2002). "Compatibilist Views on Freedom and Responsibility," *The Oxford Handbook of Free Will*, ed. Robert Kane. (2002 edition)

Haji, Ishtiyaque and Stefan Cuypers. (2008). *Moral Responsibility, Authenticity, and Education*. New York: Routledge.

Kane, Robert. (2005). *A Contemporary Introduction to Free Will*. Oxford: Oxford University Press.

Johnston, Mark. (1989). 'Dispositional Theories of Value.' *Proceedings of the Aristotelian Society* (suppl. vol.) 63, pp. 139–74.

López de Sa, Dan. (2013). 'Rigid vs Flexible Response-Dependent Properties' in Hoeltje, Schnieder & Steinberg (eds.), *Dependence*, Philosophia Verlag.

McKenna, Michael. (1998). "The Limits of Evil and the Role of Moral Address: A Defense

of Strawsonian Compatibilism," *The Journal of Ethics* 2: 123 – 142., reprinted in McKenna and Russell, eds. 2008.

__ (2012). *Conversation and Responsibility*. Oxford: Oxford University Press.

McKenna, Michael and Paul Russell, eds. (2008). *Free Will and Reactive Attitudes: Perspectives on P.F. Strawson's "Freedom and Resentment"*. Burlington: Ashgate.

Nelkin, Dana. (2011). *Making Sense of Freedom and Responsibility*. Oxford: Oxford University Press.

Pereboom, Derk. (2001). *Living Without Free Will*. Cambridge: Cambridge University Press.

__ (2013). "Free Will," in *The Oxford Handbook of the History of Ethics*.

__ (2014). *Free Will, Agency, and Meaning in Life*. Oxford: Oxford University Press.

Rosen, Gideon. (2002). "The Case for Incompatibilism," *Philosophy and Phenomenological Research* 114: 699 – 706.

Russell, Paul. (1992). "Strawson's Way of Naturalizing Responsibility," *Ethics* 102: 287 – 302.

Russell, Paul. (2011). "Moral Sense and the Foundations of Responsibility," in Robert Kane, ed. *The Oxford Handbook of Free Will*, 2011 edition, pp. 199 – 220.

Sher, George. (2006). *In Praise of Blame*. Oxford: Oxford University Press.

Strawson, P.F. (1962). "Freedom and Resentment," *Proceedings of the British Academy* 48, 1–25.

Timpe, Kevin. (2012). *Free Will: Sourcehood and Its Alternatives*. Continuum Press.

Tognazzini, Neal. (2013). "Blameworthiness and the Affective Account of Blame," *Philosophia* 41:1299–1312.

Tognazzini, Neal. (2014). "Reactive Attitudes and Volitional Necessity," *Journal of Value Inquiry* 48: 677 – 689.

Wallace, R. J. (1994). *Responsibility and the Moral Sentiments*. Cambridge: Harvard University Press.

Watson, Gary. (1987). "Responsibility and the Limits of Evil: Variations on a Strawsonian Theme." In Schoeman, Ferdinand, ed. 1987. *Responsibility, Character, and the Emotions: New Essays in Moral Psychology*. Cambridge: Cambridge University Press, 256-86., reprinted in Watson ed. 2004.

__ (2004). *Agency and Answerability: Selected Essays*. Oxford: Oxford University Press.

__ (2014). "Peter Strawson on Responsibility and Sociality," *Oxford Studies in Agency and Responsibility volume 2*, eds. David Shoemaker and Neal Tognazzini. Oxford: Oxford

University Press.

Wedgwood, Ralph. (1997). "The Essence of Response-Dependence," *European Review of Philosophy* 3: 31-54.

Wright, Crispin. (2003). *Saving the Differences: Essays on Themes from* Truth and Objectivity. Cambridge: Harvard University Press.